02 | 如何量化分析语音信号？

2021-11-24 冯建元

《搞定音频技术》 课程介绍 >

讲述：冯建元
时长 16:53 大小 15.47M

你好，我是建元。

上一讲我们了解了音频信号的一些基础知识。因为语音和音乐是最常见的音频信号类型，所以接下来两节课我会分别从这两种类型入手，带你看看如何科学、量化地对音频信号进行分析，从而读懂音频信号所承载的内容和信息，以及了解音频现象产生的原因是什么。这些知识对音频工程师、开发者甚至是调音师、录音师、音乐制作者而言都是很重要的基础，所谓“听其然，更知其所以然”。

好的，那我们这节课就先从语音信号的分析开始说起吧。

语音的基本特征

语音按照发音原理可以分为清音和浊音，语音的音调、能量分布等信息可以用基频、谐波、共振峰等特征来分析。为了更好地分析语音，我们先来看看语音是如何产生的？

浊音和清音

我们可以结合图 1 的人体发音器官结构示意图来看一下我们的语音是如何产生的。

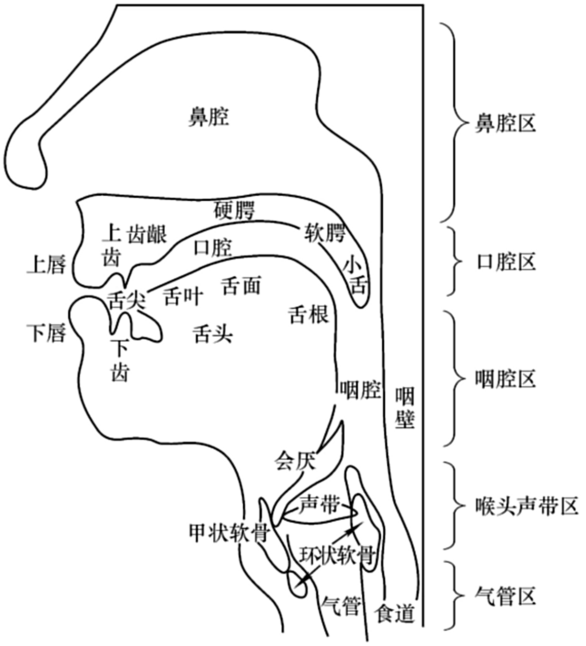


图1 人体发音器官结构示意图
新闻传播学大辞典：中国大百科全书出版社

我们知道，声道就是声音传播所通过的地方。发音的声道主要是指我们的三个腔体，即咽腔、口腔和鼻腔。而语音是由声源和声道共同作用产生的。按照声源的不同我们把语音分成以下两类：

第一类是**声带振动作为声源产生的声音，我们把它们叫做浊音**。比如拼音中的 “a，o，e” 等。

第二类是**由气体在经过唇齿等狭小区域由于空气与腔体摩擦而产生的声音，我们把它们叫做清音。**比如拼音中的 “shi、chi、xi” 等。

那么清音和浊音的声源不同在频谱上有什么样的差异呢？我们先来看看语音的频谱图（如图 2 所示）。

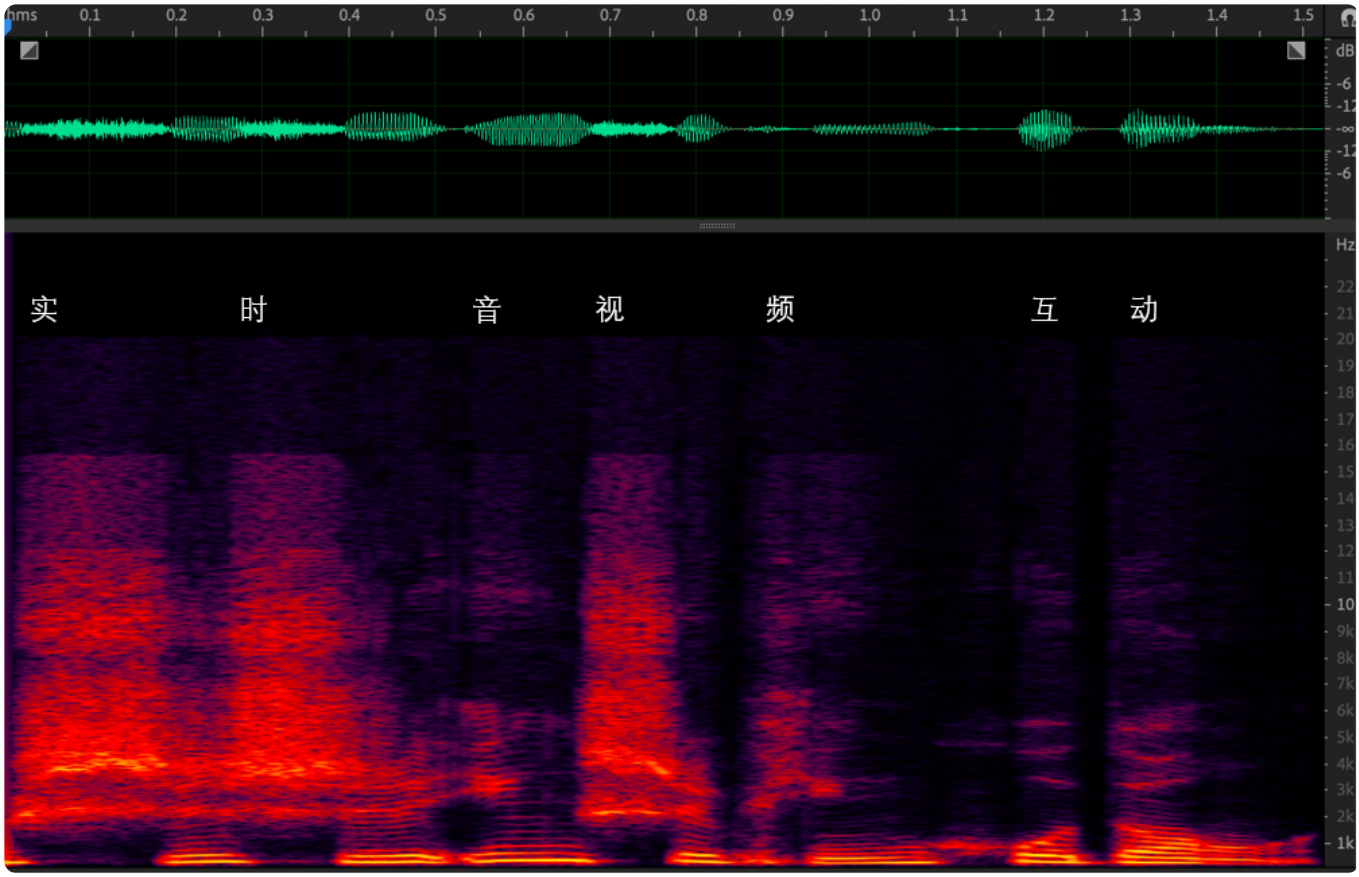


图2 语音的频谱图时域图(上)和频域图(下)

图 2 中显示的是“实时音视频互动”这几个字的音频信号的时域图和频域图（频谱图）。**时域就是信号幅度和时间的关系，而频域指的是能量与时间和频率的关系。**

频域更方便我们观察不同频率的能量分布。我们可以看到浊音，比如最后两个字“互动”是明显的有规律的能量分布，即低频能量大于高频且有明显的能量比较集中的地方，如频谱图中的亮线。而“实时”和“视”这几个字，都有 “sh” 这个由牙齿间高速气流产生的清音。清音在频谱上显示为比较均匀的分布。在 13kHz 的高频，清音也有不少的能量。

根据这个简单的分布规律我们已经可以**从频谱上分辨清浊音**了。接下来我们再来看看，我们还能从有明显能量分布的浊音的频谱中得到哪些信息。

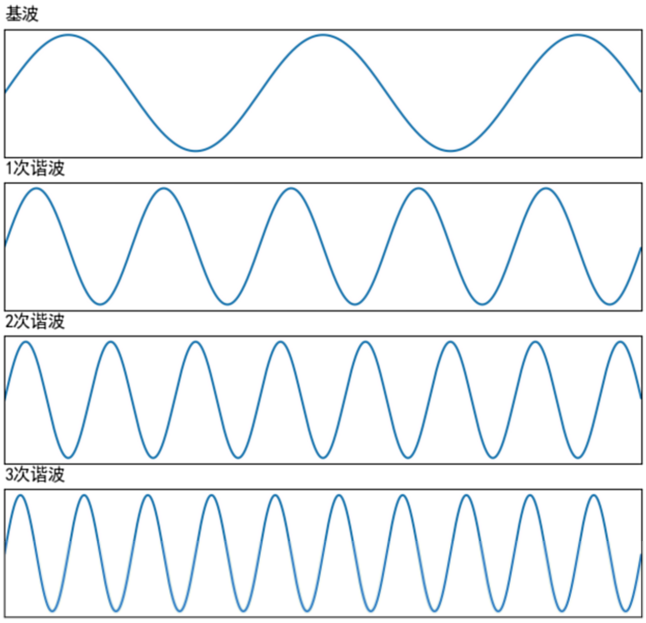
基频

在发浊音的时候，声带会振动从而产生一个声波，我们把这个声波叫做基波，并且把**基波的频率叫做基频**（一般用 F0 来表示）。**这个基频其实就可以对应到我们平时所说的音调。**比如，你唱歌音调比较高，其实就是你的声音基频比较高。

一般来说，男生的正常说话基频在 100～200Hz 之间，而女生的则会高一些，在 140～240Hz 之间。这就是为什么女生的声音听起来比男生的尖锐一些。基频会随年龄变化而变化，比如小孩的基频比较高，可以达到 300Hz，而年龄越大则基频会越来越低。基频的能量对应的是浊音频谱中频率最低的亮线。

谐波

声带振动产生的基波，在传输过程中会在声道表面反复碰撞反射，从而产生许多频率倍数于基频的声波，我们通常把这些声波叫做谐波。按照谐波频率从低到高，我们依次叫 1 次谐波，2 次谐波等等。图 3 中我们可以看一下基频信号和谐波信号在时域上的样子。



极客时间

图3 基波与谐波

谐波频率和基频是浊音能量集中的地方，这也就是为什么我们能看见浊音的频谱是一个栅格的形状。

共振峰

一个 200Hz 基频的浊音，大部分的能量都分布在 200Hz 以及 200Hz 的整数倍的频率上。那么是什么决定了哪个谐波的能量高、哪个谐波的能量低呢？

由于高次谐波是由低次谐波在腔体表面碰撞反射得到的，并且碰撞反射会导致能量的衰减，但我们在看频谱图的时候发现谐波信号并不是从低到高依次衰减的。这是为什么呢？

这是因为在这个浊音的产生过程中，声源的振动信号通过声道时，声道本身也会发生共鸣，与声道共振频率相近的能量会被增强，远离声道共振频率的部分则会被衰减，从而谐波的能量就组成了一组高低起伏的形状包络，**我们把这些包络中的巅峰位置叫做共振峰。**

比如图 4 中英文单词 father 中的 “a” 这个音我们可以看到明显的三个共振峰，频率分别为 750Hz、1100Hz、2600Hz。

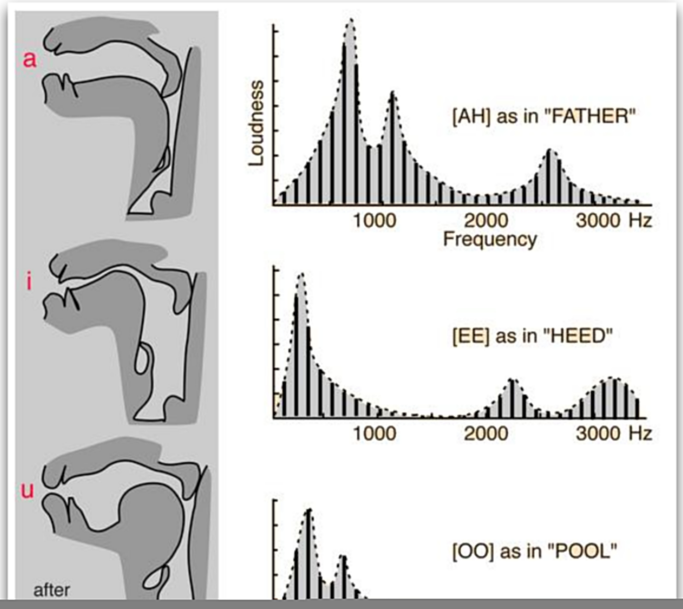


图4 共振峰示意图
语音学：标音、产生、声学和感知，中国社会科学出版社

频率从低到高我们分别用 F1、F2、F3 等来表示第一共振峰、第二共振峰、第三共振峰等。

从图 4 我们可以看到，发不同的音，比如 “a、i、u” 等，共振峰的位置和峰值都是不一样的。这是因为我们之前说的声道的三个腔体随发音的不同，开合、形状都会发生变化。从而形成了不同的腔体共振频率。所以，共振峰的位置和幅度就和发音可以一一对应起来了。这其实也是语音识别背后的原理之一，即**通过共振峰的位置和能量分布来识别音频代表的语音。**

语音信号的分析

好的，我们现在对语音是怎么产生的已经能够理解了，那接下来我们分别从时域、频域这两个方面来介绍几个常用的语音分析的方法。因为窗函数常作为时域或频域实时分析的前

处理步骤，所以在介绍这些语音分析方法之前，我们先介绍一下窗函数。

窗函数

我们分析音频时域或频域特征随时间的变化时，需要按照时间把音频截断成一个个小片段，每个小片段也就是我们说的音频帧。比如 10ms 的切片为 1 帧。

但如果直接截断信号则会导致频谱泄漏，即出现不该有的频谱分量。比如，你对一个 50Hz 的单频信号直接截断，可能会出现 60Hz、200Hz 的能量分量。因此，**我们一般采用加窗，即在原有信号中乘一个两端为 0 的窗信号，来减少截断信号时的频谱泄漏。**常用的窗函数有 Haning（汉宁窗）、Hamming（汉明窗）、Blackman（布莱克曼窗）等。在时域上加窗（Haning）的过程如图 5 所示：

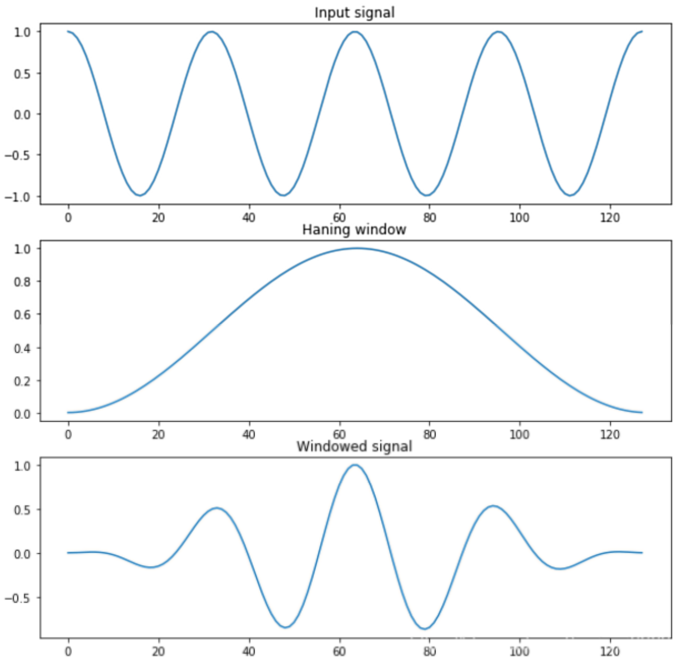


图5 时域加窗过程

可以看到图 5 中加窗的过程其实就是输入信号乘以窗函数，得到了一个两边小、中间高的新信号。

时域分析

好了，我们已经了解了窗函数，现在我们来聊聊时域分析。在时域上我们主要介绍两个指标，短时能量和短时平均过零率。

短时能量

由于语音的能量随时间的变化较快，比如能量小的时候可能就是没有在说话，而能量大的地方可能是语音中重读的地方。因此，短时能量常被用来判断语音的起止位置或者韵律。短时能量分析的定义如公式 1 所示：

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2$$

其中，x 代表采样点，w 代表窗函数。**第 n 个点的短时能量 E_n 就是由加窗后的采样信号的平方和来表示的。**由于不涉及频谱分析，因此这里的窗可以使用简单的矩形窗。

短时能量主要有以下 3 个方面的应用：

1. 可以用来区分清浊音。一般来说，清音部分的能量比浊音部分的能量要小很多。
2. 可以用来区分有声段和无声段。比如，可以设置一个能量阈值作为判断该语音段是否为静音段的条件。
3. 能量的起伏在语音识别里也被用于判断韵律（比如重读音节）的特征。

短时平均过零率

短时平均过零率，顾名思义，就是每帧内信号通过零值的次数。连续的音频信号是围绕 0 值上下波动的，并且表现为音频信号正负号随时间不断切换。短时平均过零率可以通过公式 2 和 3 来计算。

$$Z_n = 1/2 \sum_{m=n}^{n+\mathrm{N}-1} |sgn[x(m)] - sgn[x(m-1)]|$$

$$sgn[x(n)] = \begin{cases} 1, & x(n) >= 0 \\ -1, & x(n) < 0 \end{cases}$$

其中，N 为一帧中包含的信号点数，sgn 为符号函数，x 为音频采样点。

如果是**正弦信号**，例如图 3 中的基频和谐波信号，它们的短时平均过零率，就是**信号的频率除以两倍的采样频率**。

短时平均过零率在一定程度上可以表示语音信号的频率信息。由于清音的频率集中的范围要高于浊音，所以浊音的过零率要低于清音，从而**我们可以初步用短时平均过零率来判断清浊音**。

除了判断清浊音。还可以**将短时能量和短时平均过零率结合起来判断语音起止点的位置**。在背景噪声较小的情况下，短时能量比较准确；但当背景噪声比较大时，短时平均过零率有较好的效果。因此，一般的音频识别系统就是通过这两个参数相结合，来判断待检测语音是否真的开始。

频域分析

上面我们讲了基于时域的两种语音分析方法，接下来我们来学习基于频域的两种语音分析方法：短时傅里叶变换和梅尔谱。

短时傅里叶变换

短时傅里叶变换（Short-time Fourier Transform）是音频频域分析最常用的方法之一，简称 STFT。那它有什么作用呢？

我们在分析音频信号时经常会使用到频谱图（例如图 2），**那你知道这个频谱图是怎么得到的吗？**

结合短时傅里叶变换的步骤（如图 6 所示），也许你就明白了：

首先，对时域信号加滑动窗，在把音频切成若干个短帧的同时，防止频谱泄漏（窗可以使用汉宁窗）。

然后，对每一帧做快速傅里叶变换（Fast Fourier Transform，简称 FFT），把时域信号转换为复数频域信号。

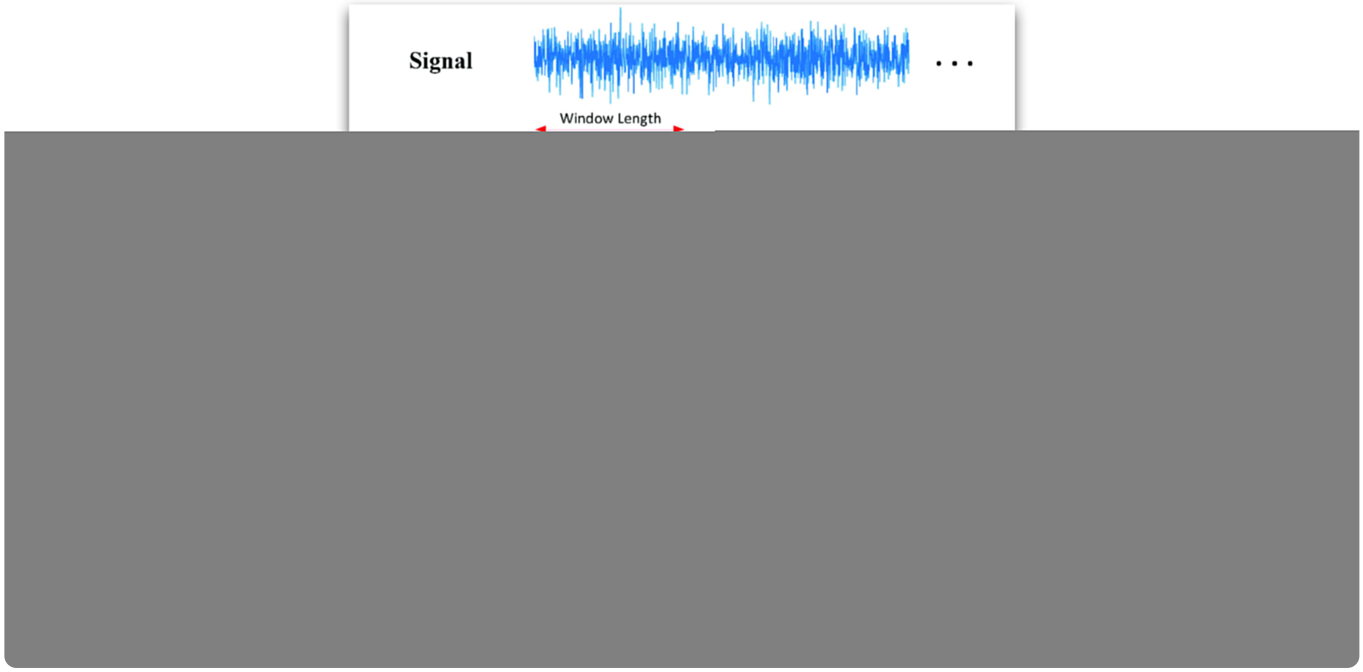


图6 短时傅里叶变换示意图
图片来源网络

图 6 中的 Hop Length 代表滑动窗移动一次的距离，并且 Overlap Length 就是两个相邻滑动窗重叠的范围。

清楚了这些，我们就可以回答刚才的问题了。其实呢，**我们是把短时傅里叶变换的结果对复数频域信号求模，并取对数转换成分贝（dB），然后用热力图的形式展示出来**，这样就能得到图 2 中的频谱图。频谱图的横坐标为时间，纵坐标为频率，并且**热力图中的颜色代表每个频点在当前时刻的能量大小**。这样我们就可以通过频谱图来观察每个时刻的语音能量分布了。

梅尔谱（Mel spectrum）

上面我们通过短时傅里叶变换得到的频谱图通常也叫做声谱、线性谱或者语谱。

由于心理和听力系统的构造，其实人耳对以 Hz 为单位的频率并不是很敏感。比如，人类很难区分 500Hz 和 510Hz 的声音。我们平时能区分的音调都是以指数排列的。比如，我们说的高八度其实就是把原有频率乘以 2。因此，用对数的频率坐标来表示可以更好地反映人的实际听感。

除此之外，人耳对不同频率声音大小的感知也是不同的。如图 7 所示，红线代表人耳感知到的响度和实际声压的对应关系，人耳感知的响度我们一般用 phon（方）来表示。

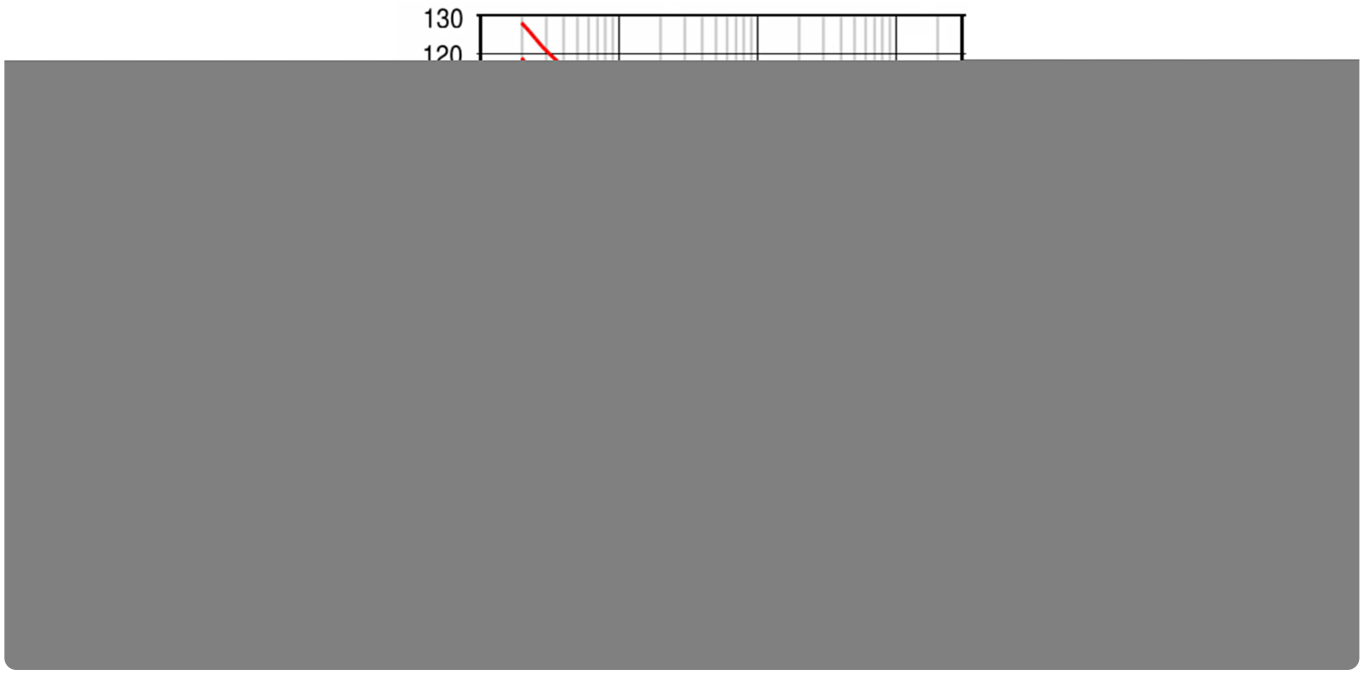


图7 人耳音量感知等响曲线
语音学：标音、产生、声学和感知，中国社会科学出版社

由图 7 可以看到，人类在 4kHz 的频率对声音的响度比较敏感，而在两端的高频和低频则需要更强的声压，人类才能感知。这其实和人类的进化有关，4kHz 多为猛兽的叫声能量分布范围，所以人耳对这类危险的频率较为警觉。

因此，为了结合人耳对频率的感知。需要使用对数的频率坐标，且通过分配滤波器组对频谱图的能量按照听感重新分配，于是就有了梅尔谱等表示形式。

Mel 谱的计算步骤分为下面几步：

首先，对语音信号进行预加重（平衡高低频能量）；

然后，语音信号通过 STFT 得到频率谱；

最后，通过三角滤波器组对频率谱逐帧进行滤波。

三角滤波器组如图 8 所示。我们可以看到三角滤波器组把频率划分成了若干个频段。敏感的频段滤波器分布比较密集，而不敏感的频段比较稀疏，这样就能更好地表征人耳的实际听感。

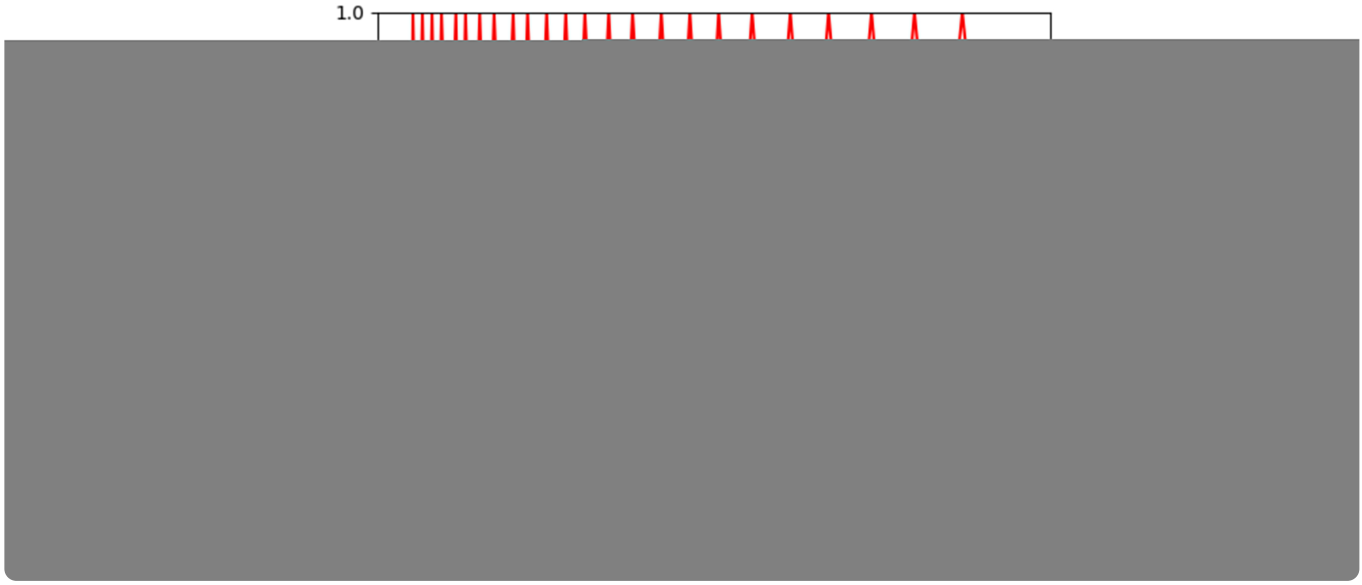


图8 梅尔谱中的三角滤波器

梅尔谱以及对梅尔谱再进一步求倒谱系数得到的 MFCC（梅尔倒谱系数），经常被用于语音识别、声音事件识别等领域。其实类似的基于人耳实际听感的表示还有 Bark 谱、Gamma Tone Filter 等，这里就不一一赘述了。如果你有兴趣可以自行了解一下，并且有问题也可以发表在留言区。

小结

好了，这节课到这里就要结束了。我们先来回顾一下这节课的重点。

关于语音是如何产生的、在信号层面有何不同、语音是如何被人耳感知的这些都是这节课的重点。有了扎实的理论基础和分析方法，以后你再碰到语音算法设计、语音特征分析时就可以有针对性地进行选择了。

现在我给你简单总结一下，我们今天主要讲了哪些内容。

1. 语音根据发音原理的不同可分为清音和浊音，并且根据它们在频谱图上的分布规律，我们可以从频谱上分辨清浊音。
2. 只有浊音才有基频和谐波。我们将发浊音时声带振动产生的声音叫基波，并且将基波的频率叫做基频。基频对应我们平时所说的音调，而谐波是频率倍数于基频的声波。
3. 共振峰表示发音受腔体形状影响。共振峰的位置和幅度可以和发音一一对应，从而我们可以通过共振峰的位置和能量分布来识别音频代表的语音。

4. 在做音频分析的时候需要对音频信号进行截断，而这会导致频谱泄漏。加窗可防止频谱泄漏。
5. 短时能量和短时平均过零率是时域分析的常见指标。它们可用于判断清、浊音以及语音的起止位置。
6. 频域分析常使用短时傅里叶变换和梅尔谱等方法，并且梅尔谱更能反映人耳的实际听感。

思考题

最后，给你介绍一个常用的 Python 音频处理工具 librosa，利用它可以快速地画出频谱图和梅尔谱。下面是示例代码，你可以尝试用自己的音频绘制一下频谱图和梅尔图，分析一下清、浊音在两个频谱上有什么不同。

复制代码

```
#绘制STFT
import numpy as np
import librosa
import matplotlib.pyplot as plt
audio,sr=librosa.load('Path',sr=48000)
n_fft=1024
ft = np.abs(librosa.stft(y[:n_fft], hop_length = n_fft+1))
plt.plot(ft)
plt.title('Spectrum')
plt.xlabel('Frequency Bin')
plt.ylabel('Amplitude')
#绘制梅尔频谱
mel_spect = librosa.feature.melspectrogram(y=y, sr=sr, n_fft=2048, hop_length=
mel_spect = librosa.power_to_db(spect, ref=np.max)librosa.display.specshow(mel
plt.title('Mel Spectrogram');
plt.colorbar(format='%+2.0f dB');
```

好了，这节课就到这里。你可以把你的问题和答案写下来，分享到留言区，与我一起讨论。我们下节课再见。

分享给需要的人，Ta订阅后你可得 20 元现金奖励

生成海报并分享

赞 1　提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇　01 | 声音是如何保存成数字信号的？

下一篇　03 | 如何分析与处理音乐信号？

精选留言 (10)

写留言

newzai
2021-11-24

老师，建议可以把参考资料，书籍罗列一下，作为深入扩展到了解。

作者回复: 其实文献就在文章里哦，比如《语音学：标音、产生、声学和感知》这本书就很推荐

12

！null
2021-11-24

上一节课入门，这一节课放弃

展开

3

哈珀朋友
2021-12-03

音视频应该特别适合电子和通信的人做，信号处理都是专业基础课啊

作者回复: 确实有不少是这个专业的人，毕竟通信的本身就包含了一部分的音频。但有些方面比如美声音效、空间音频可能就比较跨专业了。

罗进
2021-12-02

识别人声
有无及起止：短时能量

内容：过0率，浊音小
单元音：共振峰
说话人：梅尔普

展开

Geek_6a3a96
2021-11-29

老师，可以大概和我讲讲市面上k歌评分软件的道理么，我的理解是拿到一句唱句，然后做FFT获取到频谱，然后拿原唱做同样的事情，两者对比。但是我不清楚里面的更细节点。

作者回复: K歌软件打分主要是用的基频对比，既你和原唱的音调在同一时间点上是否一致。然后有的还会结合你唱歌的短时能量，比如你唱的声音是否时强时弱等来综合评断。一般不会直接用频谱，因为每个人的音色是不一样的，直接用频谱打分过于严苛了。

ZP.ZHU
2021-11-28

老师，咨询下，采用这些滤波器组处理，会不会引入很大时延？

作者回复: 在降噪中主要是一些常见的陷波滤波器，FFT滤波器等。引入的延迟并不大

ForwardsHao
2021-11-24

老师，“在背景噪声较小的情况下，短时能量比较准确；但当背景噪声比较大时，短时平均过零率有较好的效果。”这是因为背景噪声多是浊音的缘故吗？

展开

作者回复: 背景噪声只有浊音可能短时平均过零率也不好用，背景噪声一般为混合噪声过零率会大于人声中的浊音（这个假设成立的概率较大）。

ming
2021-11-24

通过对声音的时域和频域分析，可以得到什么信息？得到这些信息后，有什么用呢？

作者回复: 这节课基本上已经可以做语音的可视化了，基础主要是为了后续的音频算法设计和音频处理打好基础。

晓龙
2021-11-24

我用的是Python 3.9.7

老师的代码有一丢丢错误，改写后如下：

...
展开

作者回复: 多谢分享，不同的库版本可能有不同的依赖。

ForwardsHao
2021-11-24

这节信息量不小，得好好消化
展开

作者回复: 加油!